Integrating Multi-omics Datasets

Michael Komodromos

February 22, 2021

Table of contents

1. Methods
   - Sparse Canonical Correlation Analysis: Overview
2. Results
   - Data
   - Variable selection
   - Canonical Vectors
   - Models
   - Networks
3. Conclusion
4. Supplementary Material

Sparse Canonical Correlation Analysis: Overview

- Sparse CCA (sCCA) involves finding linear combinations of random variables that are maximally correlated.

- We can use it to understand the shared structure between datasets

sCCA: Objective Function

Let $X_1 \in \mathbb{R}^{p_1}$ and $X_2 \in \mathbb{R}^{p_2}$ be random vectors, then sCCA involves solving

$$\begin{aligned} &\text{maximise corr} \left(w_1^\top X_1, w_2^\top X_2 \right) - \lambda_1 \| w_1 \|_1 - \lambda_2 \| w_2 \|_1 \\ &\text{subject to var} \left(w_1^\top X_1 \right) \leq 1, \ \text{var} \left(w_2^\top X_2 \right) \leq 1, \end{aligned} \tag{1}$$

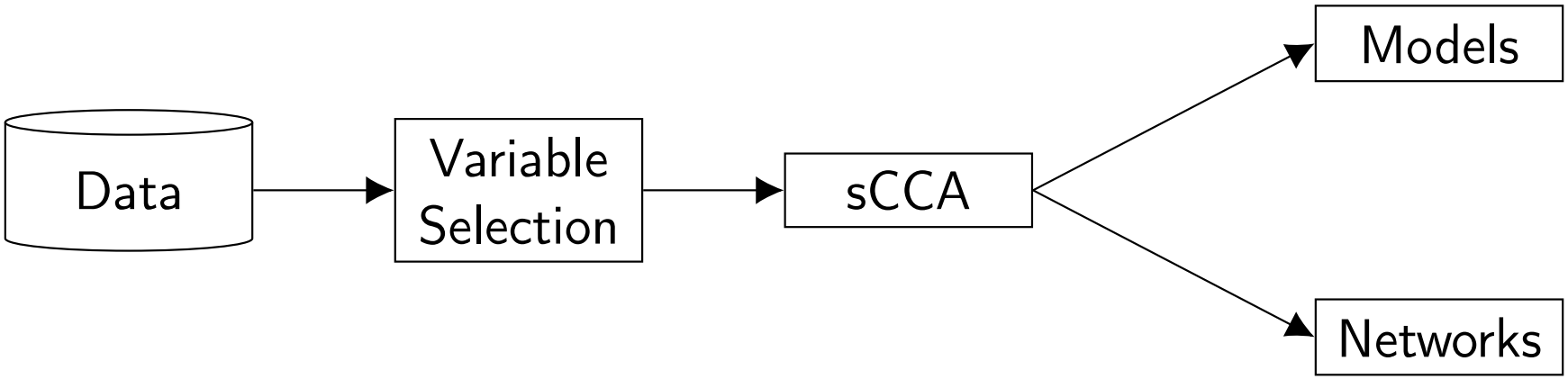
where $\lambda_1, \lambda_2 > 0$.

See: (Suo, 2018)

sCCA: Terminology

- **Canonical vectors**: $\left(w_1^{(i)}, w_2^{(i)}\right)_{i=1}^{d}$ where $1 \leq d \leq \min(p_1, p_2)$
- **Canonical variates**: $w_1^\top X_1, w_2^\top X_2$

Process overview



Data

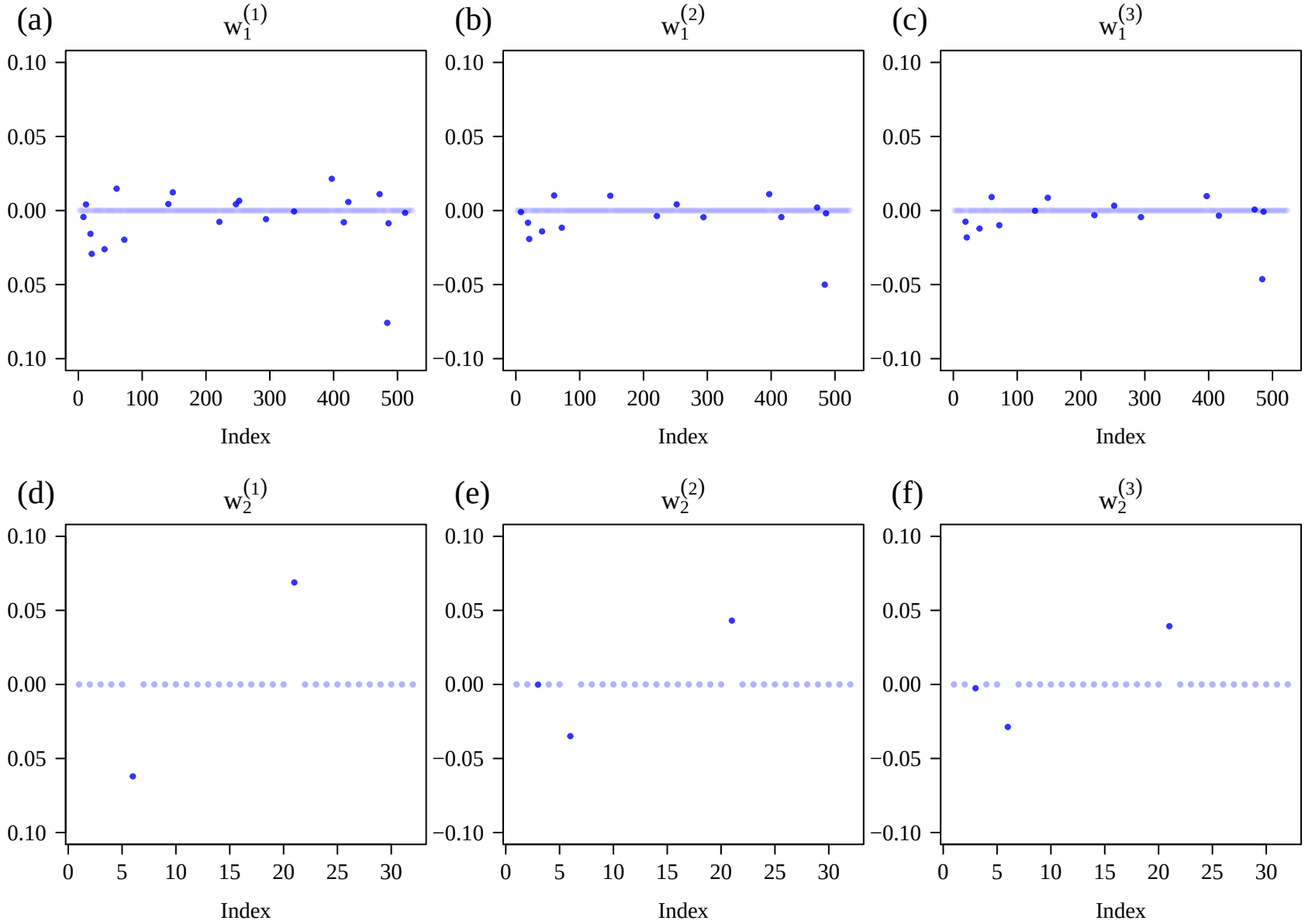
Data type	Platform	Samples	Features
Clinical		630	19
Radiomics	TexLab 2.0	71	658
mRNA expression	Affymetrix U133	593	12,043

Both the radiomics and mRNA data have more features than we have samples.

Results: Variable selection

- Pre-selection of features using a univariate cox model
- We set a threshold $\alpha = 0.05$
- $p_1 = 524$ features in our mRNA expression dataset
- $p_2 = 32$ features in our radiomics dataset

Results: Canonical Vectors



Results: Canonical Vectors

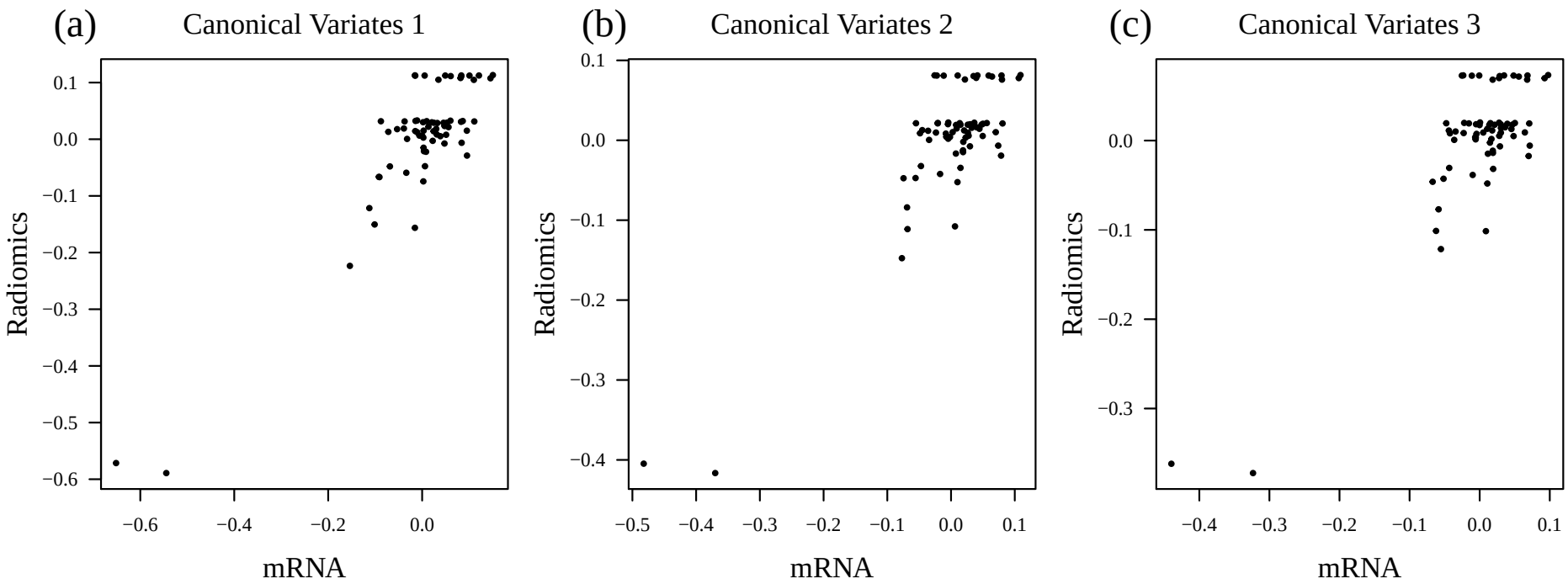
The non-zero elements in the radiomics canonical vectors are given by

- `NGTDM_Coarse_LHL_25HUgl`
- `FD_max_LLH_25HUgl`,

And the second and third canonical vectors include

- `NGTDM_Coarse_LLH_25HUgl`.

Results: Canonical Variates



Evaluating Model Performance

We use concordance as a measure of model performance.

- $\widehat{c}$ denotes Harrell's c-index and
- $\widehat{k}$ denotes a robust (to censoring) alternative.

Notably 50/68 samples are right censored.

Models: Baseline

Other Models			
Label	Predictors	$\widehat{c}$	$\widehat{k}$
M0	Age, Stage, RPV	0.672 (0.098)	0.707 (0.051)
M1	Age, Stage, RNA_1, $Radiomics_2$	0.764 (0.072)	0.761 (0.048)

Where RNA_1 and $Radiomics_2$ refers to the first and second canonical variates resp.

Models: CCA

Canonical Variates				
	Radiomics		mRNA	
Projections	$\widehat{c}$	$\widehat{k}$	$\widehat{c}$	$\widehat{k}$
1	0.631	0.709	0.713	0.716
	(0.100)	(0.055)	(0.099)	(0.050)
2	0.631	0.710	0.733	0.715
	(0.100)	(0.055)	(0.099)	(0.047)
3	0.631	0.710	0.735	0.715
	(0.100)	(0.055)	(0.099)	(0.054)

All models also include Age and Stage

Models: Principal Component Regression

Principal Components						
	Radiomics		mRNA		Radiomics, mRNA	
Projections	$\widehat{c}$	$\widehat{k}$	$\widehat{c}$	$\widehat{k}$	$\widehat{c}$	$\widehat{k}$
1	0.655	0.719	0.860	0.795	0.855	0.796
	(0.102)	(0.049)	(0.046)	(0.039)	(0.044)	(0.039)
1, 2	0.660	0.721	0.860	0.800	0.855	0.814
	(0.102)	(0.055)	(0.046)	(0.038)	(0.050)	(0.039)
1, 2, 3	0.733	0.737	0.858	0.812	0.875	0.826
	(0.065)	(0.054)	(0.049)	(0.041)	(0.049)	(0.039)

All models also include Age and Stage

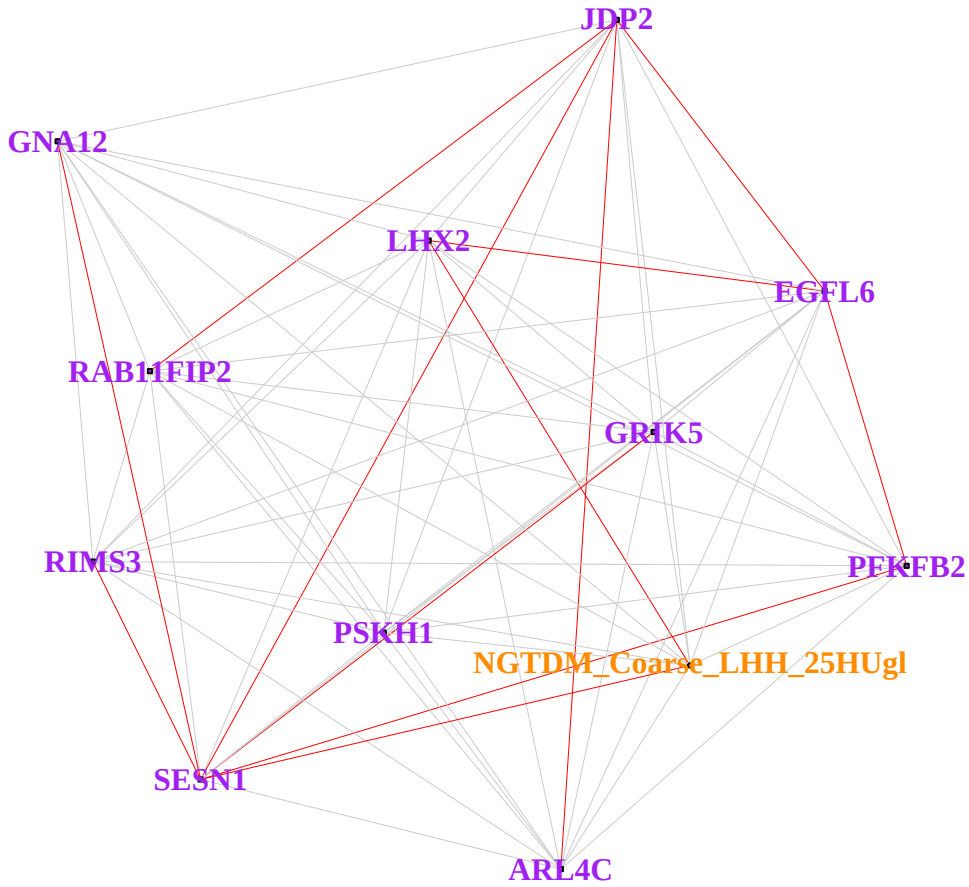
Models: Conclusions

- There is some shared structure between the radiomics and mRNA expression datasets, enough to provide models as good / better than RPV based models

- PCA based models are better than CCA based models, i.e. there is more explanatory structure within the datasets than between.

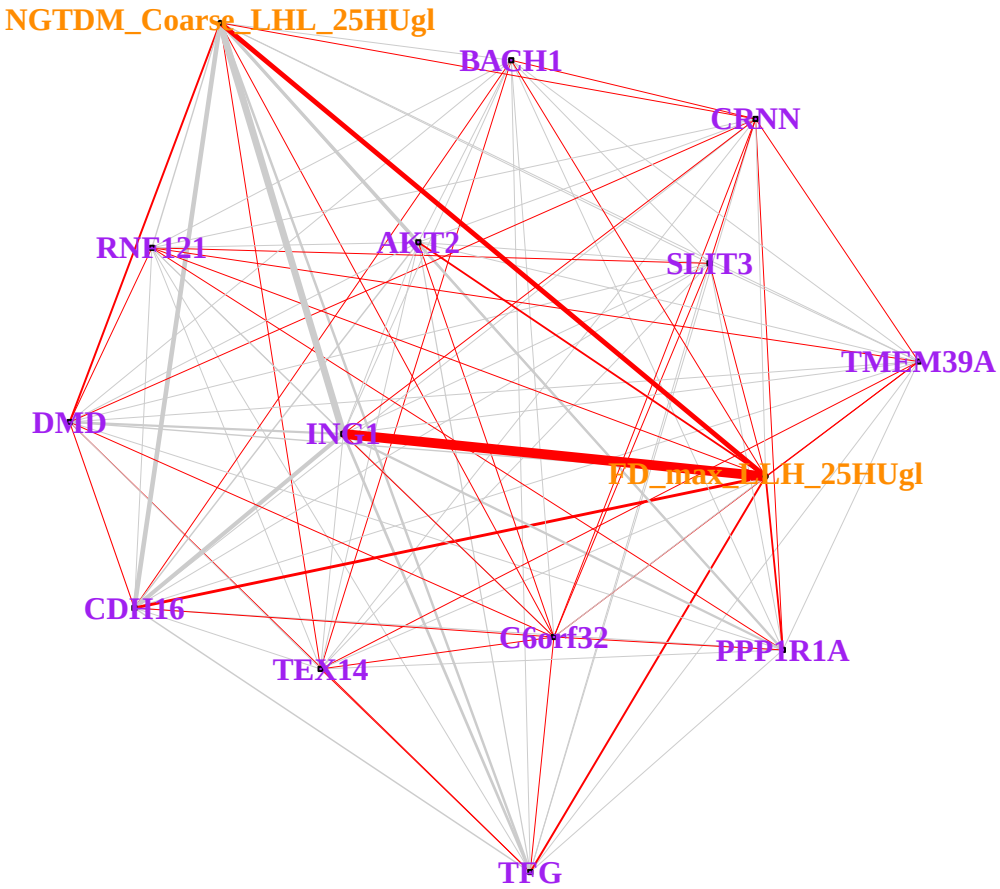
Networks

(a)


JDP2
GNA12
LHX2
EGFL6
RAB11FIP2
GRIK5
RIMS3
PFKFB2
PSKH1
NGTDM_Coarse_LHH_25HUgl
SESN1
ARL4C

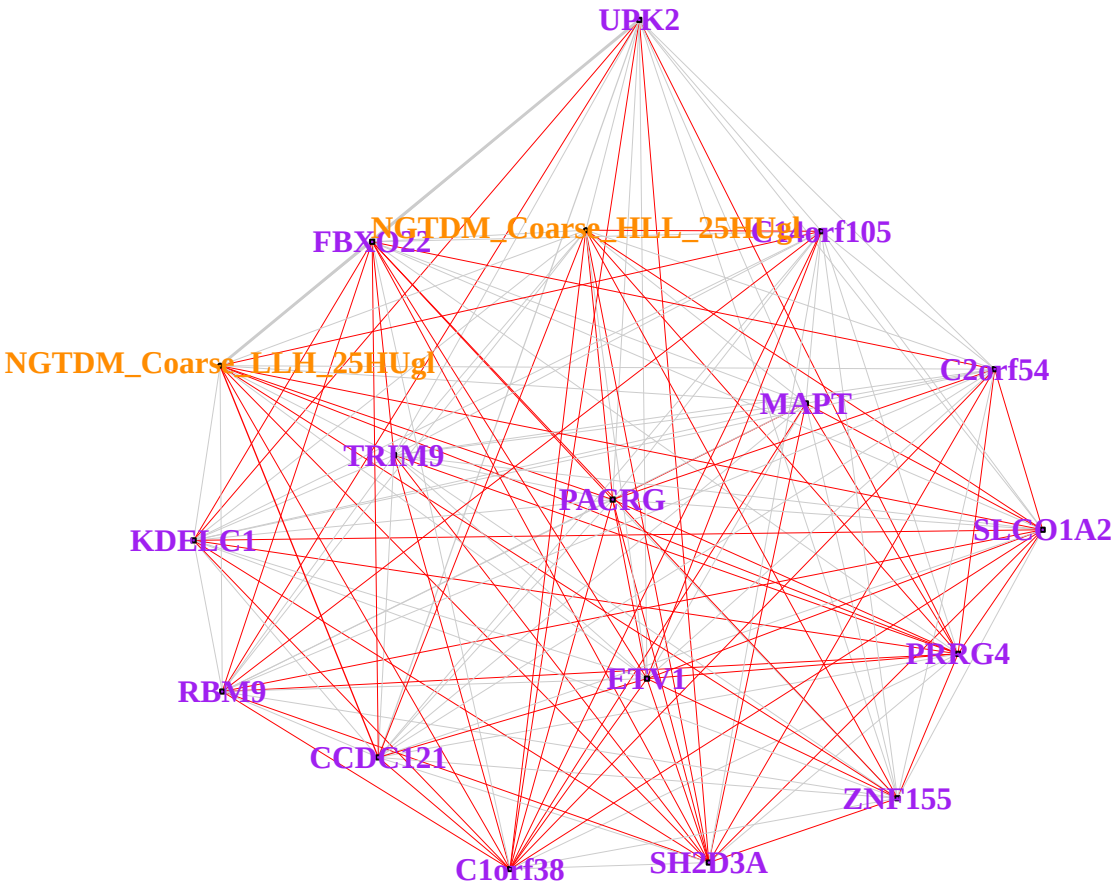
Networks

(b)


NGTDM_Coarse_LHL_25HUgl
BACH1
CBNN
RNF121
AKT2
SLIT3
TMEM39A
DMD
ING1
FD_max_LHL_25HUgl
CDH16
C6orf32
PPP1R1A
TEX14
TFG

Networks

(c)


UBK2
FBXO22
NGTDM_Coarse_HLL_25HUgl
C2orf105
NGTDM_Coarse_LLH_25HUgl
C2orf54
MAPT
TRIM9
PACRG
KDELC1
SLCO1A2
PRRG4
ETV1
RBM9
CCDC121
ZNF155
C1orf38
SH2D3A

Conclusion

- There is some structure shared between the radiomic and RNA-seq datasets
- Networks provide plausible relational pathways between mRNA and radiomics datasets
- CCA based models are as good / better than RPV based models
- mRNA PC based models are better than CCA and RPV based models

Supplementary Material: Tuning sCCA regularisation parameters

We use permutation based validation to tune the hyperparameters.

For $(\lambda_1, \lambda_2)_j \in \Lambda$

Compute (w_1^*, w_2^*) for X_1 and X_2 by solving (1) using $(\lambda_1, \lambda_2)_j$

Compute $d_j = \text{corr}(X_1 w_1^*, X_2 w_2^*)$

For $i \in \{1, \ldots, B\}$

Permute the rows of X_1 denote this matrix as X_1'.

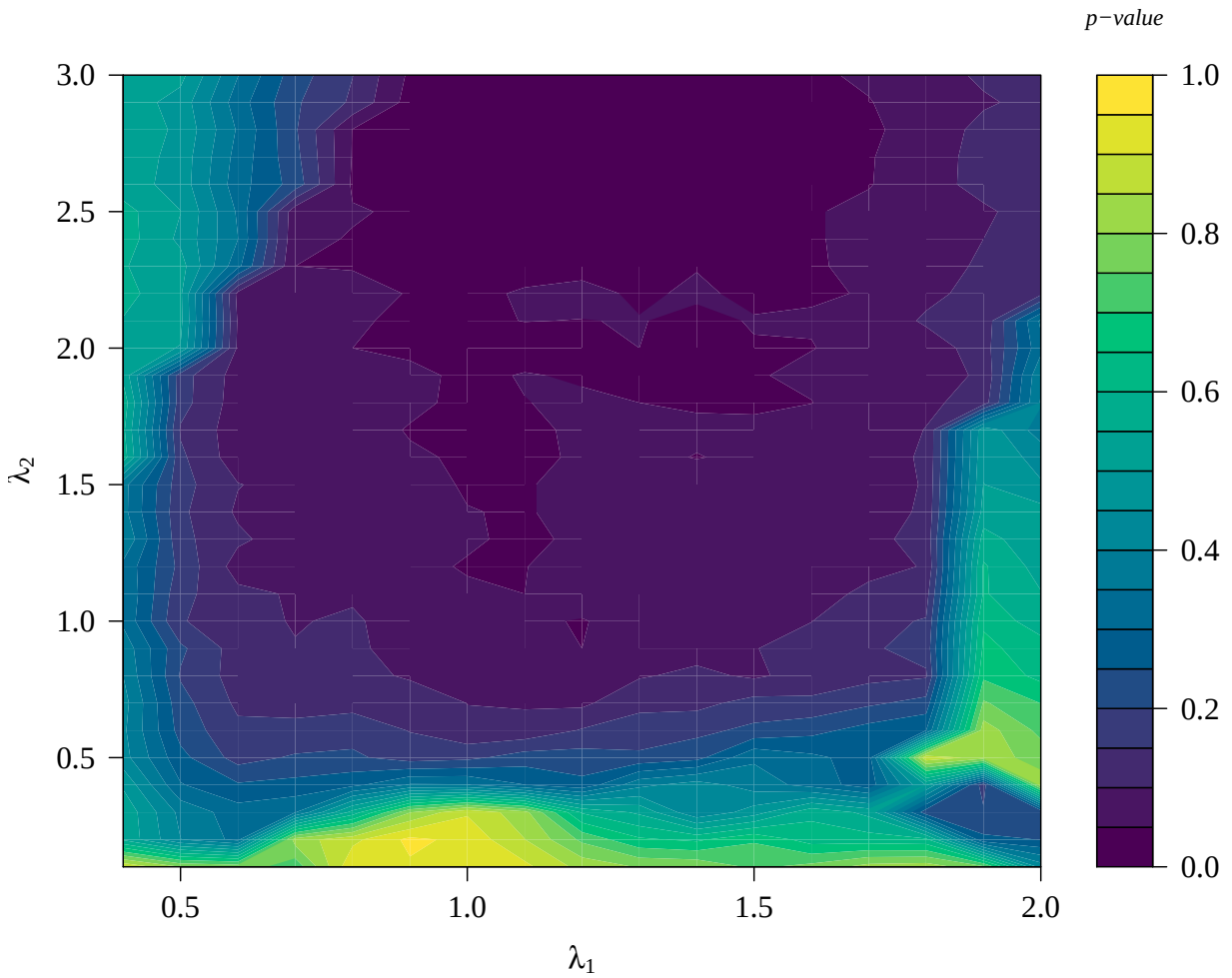
Compute (w_1', w_2') for X_1' and X_2 by solving (1) using $(\lambda_1, \lambda_2)_j$.

Compute $d_i = \text{corr}(X_1 w_1', X_2 w_2')$

Return $(\lambda_1, \lambda_2)_j$ that minimises $p_j = \frac{1}{B} \sum_{i=1}^{B} \mathbb{I}(d_i \geq d_j)$

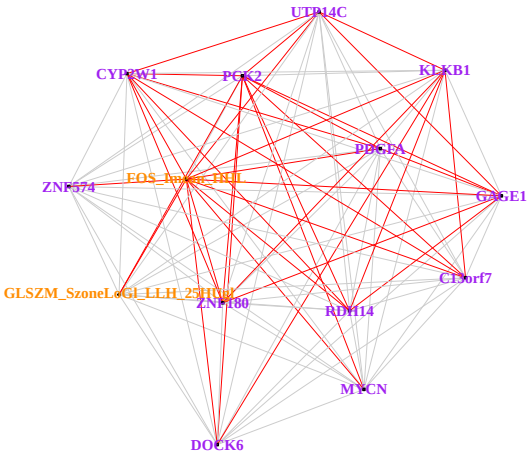
Supplementary Material: Tuning sCCA regularisation parameters

We took $\Lambda = \{0.4, 0.5, \ldots, 2\} \times \{0.1, 0.2, \ldots, 3\}$ and $B = 1000$.

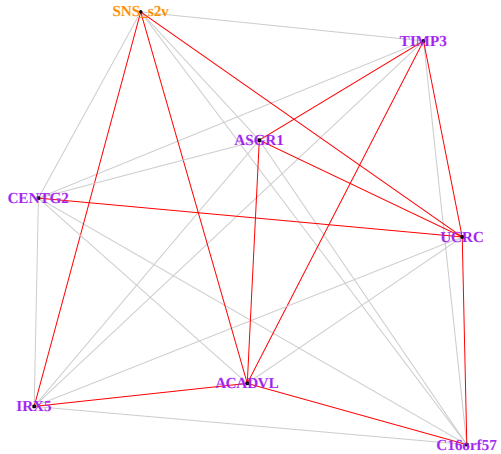


Supplementary Material: Networks

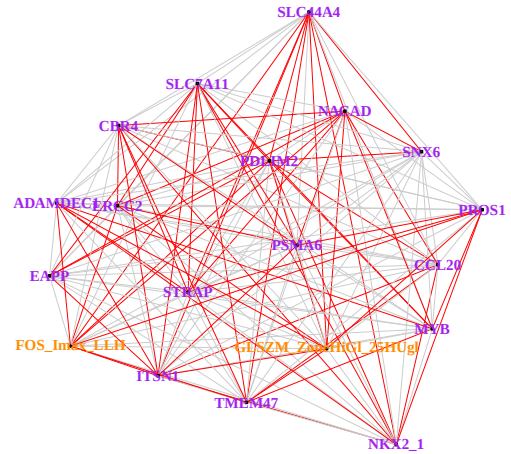
(d)



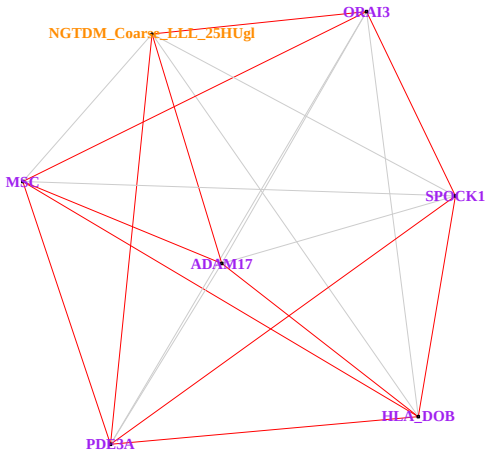
(e)



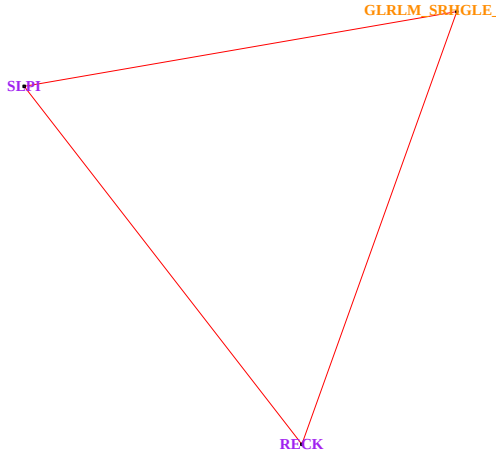
(f)



(g)



(h)

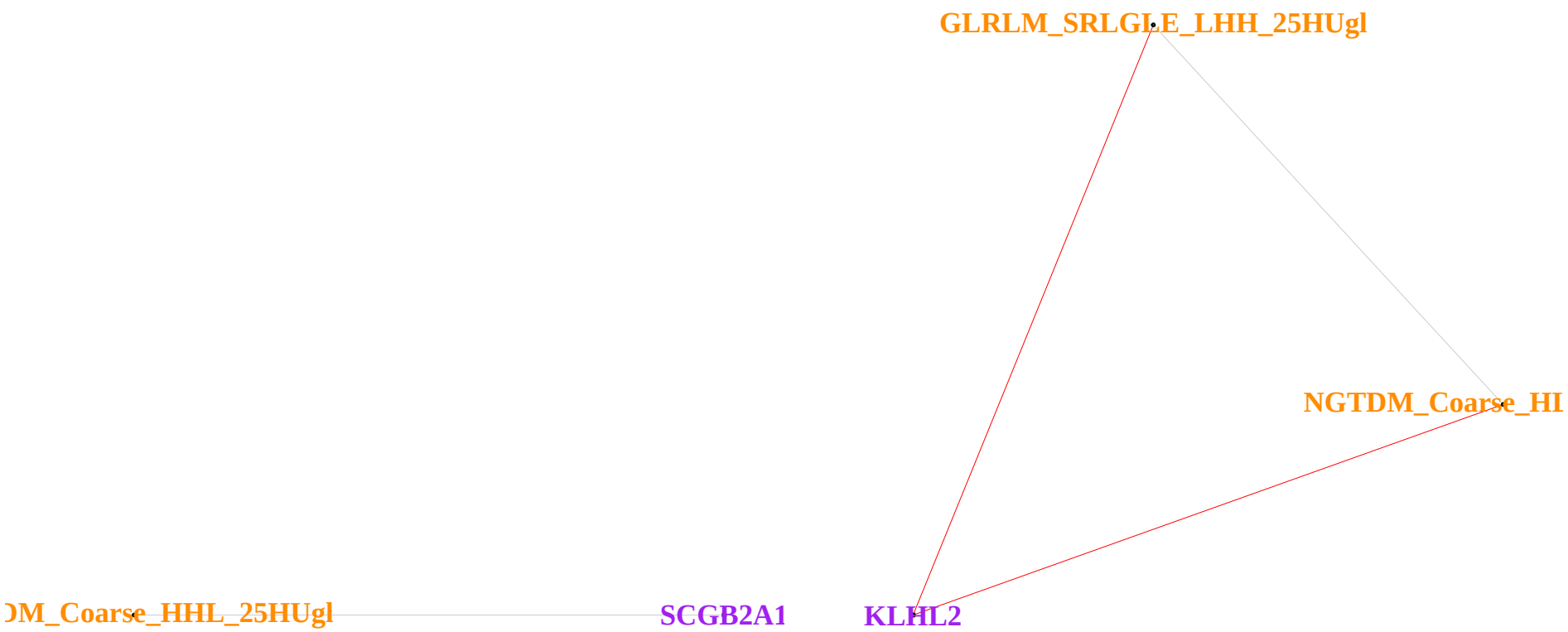


(i)

Supplementary Material: Networks

(j)

(k)



Code

This study can be reproduced by running the code at
`https://github.com/mkomod/ovc`

We also have a `R` package for sCCA available at
`https://github.com/mkomod/rcca`

References I

Gonen, M. and Heller, G. (2005). Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, 92(4):965–970.

Hotelling, H. (1936). Relations Between Two Sets of Variates. *Biometrika*, 28(3/4):321.

Rodosthenous, T., Shahrezaei, V., and Evangelou, M. (2020). Integrating multi-OMICS data through sparse canonical correlation analysis for the prediction of complex traits: a comparison study. *Bioinformatics*, 36(17):4616–4625.

Shi, W. J., Zhuang, Y., Russell, P. H., Hobbs, B. D., Parker, M. M., Castaldi, P. J., Rudra, P., Vestal, B., Hersh, C. P., Saba, L. M., and Kechris, K. (2019). Unsupervised discovery of phenotype-specific multi-omics networks. *Bioinformatics*, 35(21):4336–4343.

Suo, X. (2018). *Topics In High-Dimensional Statistical Learning*. PhD thesis, Stanford.

References II

Uurtio, V., Monteiro, J. M., Kandola, J., Shawe-Taylor, J., Fernandez-Reyes, D., and Rousu, J. (2017). A tutorial on canonical correlation methods. *arXiv*, 50(6).

Witten, D. M. and Tibshirani, R. J. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat. Appl. Genet. Mol. Biol.*, 8(1):1–27.